

# NEALT Proceedings Series Vol. 36

Proceedings of the 7th Workshop on  
NLP for Computer Assisted Language Learning (NLP4CALL 2018)

at SLTC, Stockholm, 7th November 2018





Proceedings of the  
7th Workshop on  
NLP for Computer Assisted Language Learning  
(NLP4CALL 2018)  
at SLTC 2018  
Stockholm, 7th November 2018

edited by

Ildikó Pilán, Elena Volodina, David Alfter and Lars Borin

**Front cover photo** by Stockholm University Library  
Licensed under Creative Commons Attribution 2.0 Generic:  
<https://creativecommons.org/licenses/by/2.0/>

Linköping Electronic Conference Proceedings  
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print)  
ISBN 978-91-7685-173-9

152  
2018



## Preface

The primary goal of the workshop series on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL) is to create a meeting place for researchers working on the integration of Natural Language Processing and Speech Technologies in CALL systems and exploring the theoretical and methodological issues arising in this connection. The latter includes, among others, insights from Second Language Acquisition (SLA) research, on the one hand, and promoting the development of “Computational SLA” through setting up Second Language research infrastructure(s), on the other.

The intersection of Natural Language Processing (or Language Technology / Computational Linguistics) and Speech Technology with Computer-Assisted Language Learning (CALL) brings “understanding” of language to CALL tools, thus making CALL intelligent. This fact has given the name for this area of research – Intelligent CALL, ICALL. As the definition suggests, apart from having excellent knowledge of Natural Language Processing and/or Speech Technology, ICALL researchers need good insights into second language acquisition theories and practices, as well as knowledge of second language pedagogy and didactics. This workshop invites therefore a wide range of ICALL-relevant research, including studies where NLP-enriched tools are used for testing SLA and pedagogical theories, and vice versa, where SLA theories, pedagogical practices or empirical data are modeled in ICALL tools. The NLP4CALL workshop series is aimed at bringing together competencies from these areas for sharing experiences and brainstorming around the future of the field.

### We invited submissions:

- that describe research directly aimed at ICALL;
- that demonstrate actual or discuss the potential use of existing Language and Speech Technologies or resources for language learning;
- that describe the ongoing development of resources and tools with potential usage in ICALL, either directly in interactive applications, or indirectly in materials, application or curriculum development, e.g. learning material generation, assessment of learner texts/responses, individualized learning solutions, provision of feedback;
- that discuss challenges and/or research agenda for ICALL;
- that describe empirical studies on language learner data.

A special focus was given to established and upcoming infrastructures aimed at SLA and learner corpus research, covering questions such as data collection, legal issues, reliability of annotation, annotation tool development and search environments for SLA-relevant data. We encouraged paper presentations and software demonstrations describing the above-mentioned themes primarily, but not exclusively, for the Nordic languages.

This year, we had the pleasure to welcome two invited speakers: Jill Burstein (Educational Testing Service) and Jan Hulstijn (University of Amsterdam).

**Jill Burstein** is a Research Director of the Natural Language Processing Group in Research & Development at Educational Testing Service in Princeton, New Jersey. Her research interests span Natural Language Processing for educational technology, automated essay scoring and evaluation, discourse and sentiment analysis, argumentation mining, education policy, English language learning, and writing research. The intersection of her interests has led to two extensively used commercial applications for English L2 learners: E-rater®, ETS' automated essay evaluation application, and the Language Muse Activity Palette™ - a new classroom tool under development targeting English learners that automatically generates

language activities for classroom texts to support content comprehension. Jill Burstein is one of the most successful researchers within ICALL that together with a group of bright researchers made ICALL tools a reality for many teachers of L2 English.

**In her talk**, *Natural Language Processing for Education: Applications for Reading and Writing Proficiency*, she explored automated writing evaluation (or, AWE) systems, which have been largely used to support the measurement of writing skills for on-demand, large-volume, high-stakes assessments. She argued that advances in natural language processing (NLP)-driven AWE now affords the ability, in real-time, to generate a variety of linguistic information which can provide support literacy for reading and writing. NLP-based technology can now be used to 1) build a broader array of capabilities to support the instruction for a diverse population of learners, and 2) offer educational analytics for various stakeholders, including students, instructors, parents, administrators and policy-makers. Her talk discussed the history of AWE, the literacy-based motivation and trajectory for AWE-driven technology development, two technology use cases of AWE-based reading (The Language Muse® Activity Palette) and writing applications (The Writing Mentor™), and exploratory research examining relationships between linguistic features in college writing and broader success predictors that can potentially inform continued development of technology that supports literacy.

**Jan Hulstijn** is professor emeritus of second language acquisition at the Amsterdam Center for Language and Communication (ACLC) of the University of Amsterdam. He has been affiliated with this university (full professor) since 1998. Before that he held positions at Leiden University and the Free University of Amsterdam. He was associate post-doc researcher at the University of Toronto, Canada (1982-1983) and he was visiting professor at the University of Leuven, Belgium, (2002) and at Stockholm University (2005). His main research interests are concerned with (1) language proficiency in native and non-native speakers; (2) explicit versus implicit accounts of first and second language learning and (3) theories of second language acquisition and philosophy of science. With others PIs, he received a number of research grants from the Netherlands Organisation of Scientific Research (NWO) between 1982 and 2007 (see webpage). In 2018 he received the 2018 distinguished scholar award from the European Second Language Association (EuroSLA) (<http://www.eurosla.org/distinguished-scholar-award-2018-jan-h-hulstijn/>). In 2015, he published a book presenting his theory of basic language cognition (BLC).

**His talk** was entitled *Usage-based views on second language acquisition and the Common European Framework of Reference (CEFR): their potential relevance for the NLP field*. The first wave of the Cognitive Revolution (1960 – 1985) was dominated by (1) the competence-performance distinction and Universal Grammar in generative linguistics, and by (2) notions of modularity and serial processing in psycholinguistics. These notions also dominated the scientific study of second language acquisition (SLA) at the time. In contrast, more recent work in SLA is increasingly being influenced by ideas which originated during the second wave of the Cognitive Revolution (1985 – the present), in particular usage-based linguistics, Emergentism and the Competition Model, Construction Grammar, Dynamic Systems Theory, implicit/statistical learning, and statistical learning in relation to corpus linguistics. These developments might be important for people working in the field of NLP. In the first part of his talk he gave a brief overview of these more recent developments, with a particular focus on the unified conceptualization of representation and processing (referred to together with the term cognition), the notion of graded cognition (as opposed to dichotomous views of cognition), the notion of frequency and recency of linguistic elements in learners' input (with the aid of corpus linguistics), the removal of traditional barriers between lexis and grammar, and the need to explain individual differences in language knowledge and use. In his view,

current software for automatic analysis of corpora of spoken or written language production is still incapable of identifying grammatical constructions relevant from an SLA perspective. The notions of shared/basic and non-shared/extended language cognition were also briefly introduced (Hulstijn, 2015, 2018). In the final part of his talk, he presented his views on the Common European Framework of Reference for Language (CEFR, 2001), which currently dominates almost all practices in second-language testing in Europe and whose presence has been increasing also in other parts of the world.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. Amsterdam: John Benjamins Publishing Company.

Hulstijn, J.H. (2018, early view). An individual-differences framework for comparing non-native with native speakers: Perspectives from BLC Theory. *Language Learning*. DOI: 10.1111/lang.12317

### **Previous workshops**

This workshop follows a series of workshops on NLP for CALL organized by the NEALT Special Interest Group on Intelligent Computer-Assisted Language Learning (SIG-ICALL<sup>1</sup>). The workshop series has previously been financed by the Center for Language Technology<sup>2</sup> at the University of Gothenburg, and the Swedish Research Council's conference grant.

Submissions to the seven workshop editions have targeted a wide variety of languages, ranging from well-resourced languages (Chinese, German, English, French, Portuguese, Russian, Spanish) to less-resourced ones (Erzya, Arabic, Estonian, Irish, Komi-Zyrian, Meadow Mari, Saami, Udmurt, Võro). Among these several Nordic languages have been targeted: Danish, Estonian, Finnish, Icelandic, Norwegian, Saami, Swedish, and Võro. The wide scope is also evident in the affiliations of the participating authors as shown in Table 1:

---

<sup>1</sup> <https://spraakbanken.gu.se/swe/forskning/ICALL/SIG-ICALL>

<sup>2</sup> <http://clt.gu.se/>

<b>Country</b>	<b># authors</b>
Australia	2
Belgium	4
Canada	3
Denmark	1
Estonia	3
Finland	7
France	3
Germany	57
Iceland	3
Ireland	2
Japan	2
Norway	12
Portugal	4
Russia	10
Slovakia	1
Spain	3
Sweden	62
Switzerland	10
UK	1
US	3

Table 1: Authors by affiliation country, 2012-2018

During the past years, the acceptance rate has varied between 50% and 77%, the average being 65% (see Table 2). The acceptance rate is rather high, however, the reviewing process has always been very rigorous, with two or three double-blind reviews per submission. This indicates that submissions to the workshops have usually been of high quality.

<b>Workshop year</b>	<b>Submitted</b>	<b>Accepted</b>	<b>Acceptance rate</b>
2012	12	8	67%
2013	8	4	50%
2014	13	10	77%
2015	9	6	67%
2016	14	10	72%
2017	13	7	54%
2018	16	11	69%

Table 2: Submission and acceptance rates, 2012-2018

We would like to thank our Program Committee for providing detailed feedback for the reviewed papers:

- Lars Ahrenberg, Linköping University, Sweden
- David Alfter, University of Gothenburg, Sweden
- Lisa Beinborn, University of Duisburg-Essen, Germany
- Eckhard Bick, University of Southern Denmark, Denmark
- António Branco, University of Lisbon, Portugal



- Jill Burstein, Educational Testing Services, US
- Andrew Caines, University of Cambridge, UK
- Dirk De Hertog, KU Leuven, Belgium
- Simon Dobnik, University of Gothenburg, Sweden
- Thomas François, UCLouvain, Belgium
- Johannes Graën, University of Zurich, Switzerland
- Andrea Horbach, University of Duisburg-Essen, Germany
- John Lee, City University of Hong Kong, Hong Kong
- Peter Ljunglöf, University of Gothenburg and Chalmers University of Technology, Sweden
- Montse Maritxalar, University of the Basque country, Spain
- Beáta Megyesi, Uppsala University, Sweden
- Detmar Meurers, University of Tübingen, Germany
- Martí Quixal, University of Tübingen, Germany
- Robert Reynolds, Brigham Young University, USA
- Gerold Schneider, University of Zurich, Switzerland
- Irina Temnikova, Sofia University, Bulgaria
- Francis Tyers, The Arctic University of Norway, Norway
- Sowmya Vajjala, Iowa State University, US
- Elena Volodina, University of Gothenburg, Sweden
- Mats Wirén, Stockholm University, Sweden
- Victoria Yaneva, University of Wolverhampton, UK
- Torsten Zesch, University of Duisburg-Essen, Germany
- Robert Östling, University of Helsinki, Finland

We intend to continue this workshop series, which so far has been the only ICALL-relevant recurring event based in the Nordic countries. Our intention is to co-locate the workshop series with the two major LT events in Scandinavia, SLTC (the Swedish Language Technology Conference) and NoDaLiDa (Nordic Conference on Computational Linguistics), thus making this workshop an annual event. Through this workshop, we intend to profile ICALL research in Nordic countries as well as beyond, and we aim at providing a dissemination venue for researchers active in this area.

Workshop website:

<https://spraakbanken.gu.se/eng/icall/7th-nlp4call>

*Workshop organizers*

*Ildikó Pilán, Elena Volodina, David Alfter, Lars Borin  
Språkbanken, University of Gothenburg*

## **Acknowledgements**

We gratefully acknowledge the financial support from *SweLL infrastructure project on Swedish Second language*<sup>3</sup> regarding the invited speaker Jill Burstein, and the project on *Development of lexical and grammatical competences in immigrant Swedish*<sup>4</sup>, that provided funding for our other invited speaker, Jan Hulstijn.

---

<sup>3</sup> [https://spraakbanken.gu.se/eng/swell\\_infra](https://spraakbanken.gu.se/eng/swell_infra), <https://rj.se/anslag/2016/swell---forskningsinfrastruktur-forsvenska-som-andrasprak/>

<sup>4</sup> <https://rj.se/en/anslag/2017/utveckling-av-lexikala-och-grammatiska-kompetenser-i-invandrarsvenska/>, <https://spraakbanken.gu.se/eng/12-profiling>



## Contents

Preface	i
<i>Ildikó Pilán, Elena Volodina, David Alfter and Lars Borin</i>	
Using authentic texts for grammar exercises for a minority language	1
<i>Lene Antonsen and Chiara Argese</i>	
Normalization in context: Inter-annotator agreement for meaning-based target hypothesis annotation	10
<i>Adriane Boyd</i>	
The role of diacritics in increasing the difficulty of Arabic lexical recognition tests	23
<i>Osama Hamed and Torsten Zesch</i>	
An automatic error tagger for German	32
<i>Inga Kempfert and Christine Köhn</i>	
Demonstrating the MUSTE language learning environment	41
<i>Herbert Lange and Peter Ljunglöf</i>	
Learner corpus anonymization in the age of GDPR: Insights from the creation of a learner corpus of Swedish	47
<i>Beáta Megyesi, Lena Granstedt, Sofia Johansson, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén and Elena Volodina</i>	
Work smart – Reducing effort in short-answer grading	57
<i>Margot Mieskes and Ulrike Padó</i>	
NLP Corpus Observatory – Looking for constellations in parallel corpora to improve learners’ collocational skills	69
<i>Gerold Schneider and Johannes Graÿn</i>	
A linguistically-informed search engine to identify reading material for functional illiteracy classes	79
<i>Zarah Weiss, Sabrina Dittrich and Detmar Meurers</i>	
Feedback strategies for form and meaning in a real-life language tutoring system	91
<i>Ramon Ziai, Bjoern Rudzewitz, Kordula De Kuthy, Florian Nuxoll and Detmar Meurers</i>	